INTERACTIVE TOOL FOR VISUALIZATION OF TOPIC MODELS

Miroslav SMATANA, Viktória MARTINKOVÁ, Dominika MARŠÁLEKOVÁ, Peter BUTKA
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, tel. +421 55 602 4220,
e-mail: miroslav.smatana@tuke.sk, viktoria.martinkova@student.tuke.sk, dominika.marsalekova@student.tuke.sk,
peter.butka@tuke.sk

ABSTRACT

Digital data are all around us and occurs in various forms as videos, pictures or texts. Digital documents represent the vast majority of such data. It can be e-news, social media contributions and so on. They can contain useful information, but due to their amount, it is time-consuming to find relevant information for the concrete company or persons. For that reason, there is a need for their automatic analysis. One of the areas which dealt with textual data analysis is topic modeling. It showed us a new way of how to automatically browse, search and summarize data in the organization. Topic modeling can be useful for time-based analysis of crises, elections, news feeds, launching of new products on the market, and other tasks which led to decision support tasks. In this paper, we aim to survey and compare topic modeling methods and propose web application to visualize extracted topics using topic modeling method called Latent Dirichlet Allocation (LDA). The comparison of selected standard topic modeling methods was experimentally tested on two selected textual datasets (20Newsgroup and Reuters) using standard evaluation metric. The proposed web application was implemented to use LDA and can extract topic models from textual documents datasets, visualize them and show their evolution over time.

Keywords: *topic modeling, visualization, data analysis, Latent Dirichlet Allocation*

1. INTRODUCTION

In recent days there is a lot of digital data, especially in the form of digital documents containing textual data. They occur variously as e-news, blogs, social media contributions, and so on. Those data can be beneficial for the competitiveness of organization, understanding of real-life events and processes. It means that digital documents can be searched, navigated and visualized in the way which helps people or organizations. For example, it can contain a lot of information, which can be used by companies in several decision-making situations like:

- Launching of new products to the market - this problem is related to the analysis of similar products on the market and contributions about them, their issues and needs for a new product.
- Targeted marketing – helps to focus some specific groups of customers according to their shared topics.
- Protect of reputation – the company can try to identify topics that are related to them, analyze them according to their polarity (positive/negative) and decide for some change in their strategies.
- Crisis / Risk analysis – analysis of topics in the moment of some crisis or risk mitigation needs related to current issues in the company.
- Etc.

However, the problem of data processing and analysis is in their large amount, so there is a need for new ways of their automatic analysis. For now, there exist several methods, which are suitable for automatic analysis of textual data.

In this paper, we analyze one family of these methods from the field known as topic modeling. The methods from this area can provide new approaches for searching, retrieving and summarizing information in large data corpus. Topic modeling, especially in connection with social networks, can be useful for time-based analysis of digital documents content and evolution of contributions in time. The main idea is to extract topics, which are in document or contribution, visualize them and provide their evolution (changes) during the time.

Our main aim in this paper was to describe some of the methods and compare them, to select one of the methods for a proposed web application – search and visualization platform.

The rest of the papers shows the main purpose of topic modeling as uncovering the hidden thematic structure from input documents, and show new ways to browse, search and summarize textual data. Several methods, which solve the problem of topic modeling were proposed, and we describe them in Section 2. In next Section 3, we provided experiments with standard methods for topic modeling, which can be then applied in practice.

Another problem of processing and analysis is appropriate visualization of analysis results to end user. For that reason, we propose web application to present topic modeling result in easy and intuitive form to end user. This proposed application is described in Section 4.

2. TOPIC MODELING METHODS

As was written topic modeling represent methods, which are trying to uncover hidden thematic structure from input data and it shows us a new way of searching, browsing and summarizing those data.

An example of topic modeling output (for example of an article from news in Associated Press) can be in this

form: *"The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Julliard School. Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important..."*, where every color in the text represents a different topic (for example red color - arts, green color - budgets, etc.).

There already exist several methods of topic modeling. Latent Semantic Indexing (LSI) [1] can be seen as one of the first methods which solves this problem because it tries to uncover semantic structure form text, but it is not usually introduced as standard topic modeling method. However, it became base for other topic modeling methods. One of them is Latent Dirichlet Allocation (LDA) [2]. This method belongs to the most widely used approach for topic modeling. Also, several extensions of LDA were proposed. For example, by Petterson et al. [3] or Zhai and Graber [4] present an interesting algorithm for online topic modeling, they proposed an online version of LDA.

In recent years there has also been created other methods, not only modifications of LDA. Yee et al. in their paper [5] proposed Hierarchical Dirichlet Process (HDP), Li et al. [6] proposed topic modeling based on moving average stochastic variation inference, Hofmann [7] presents method based on stochastic variational inference, Phan et al. [8] proposed method using conventional topic model based on external data, Sridhar [9] uses for topic modeling from short text Gaussian mixture models. Another interesting approach was proposed by Quan et al. [10], where they transform input short texts into long pseudo-document before the application of the standard conventional method.

Mentioned methods are used for standard topic modeling (extracting topic on one level of abstraction with no information about time), however nowadays with such amount of data, there is a need for more informative analysis. For that reason, there were created several subtasks of topic modeling which offer more detailed analysis.

First is topic modeling over time, which aim to capture topics evolution over time. In this case, Blei and Lafferty [11] present a method called Dynamic Topic Models (DTM), which belongs to the family of probabilistic time series models and provides time evolution of topics in input collections of texts, Wang et al. present DTM extension called Continuous Time Dynamic Topic Models (cDTM) [12] and Beykikhoshk et al. in [13] present a different approach to capturing topic time evolution based on the Hierarchical Dirichlet Process.

Second is hierarchical topic modeling, which try to extract topics on several levels of abstraction. Different methods have already been developed to achieve such a goal, e.g., Blei et al. presented a method based on a nested Chinese restaurant process [14], Hoffman described a cluster-based method [15], Smith et al. [16] provided the hierarchical version of LDA. From other approaches, we can mention Pachinko Allocation Model (PAM) [17], Hierarchical Latent Tree Analysis (HLTA) [18] or a method presented in [19], where authors developed a hierarchical model based on HDP.

In this paper we focus on two basic aspects work only on standard topic models.

3. COMPARISON OF STANDARD TOPIC MODELING METHODS

This part includes the evaluation of extracted topic quality by different topic models, i.e., LSI - Latent Semantic Indexing, LDA - Latent Dirichlet Allocation, and HDP - Hierarchical Dirichlet Process. First two were evaluated using standard evaluation metrics and their values through a different number of topics. HDP was evaluated only for average values of a specific number of topics extracted by method.

The evaluation was performed on the 20Newsgroups dataset[1], which contains 18846 documents divided into 20 classes, and on the Reuters Dataset[2], which contains 90 classes and 10788 documents. For all of the evaluated methods we preprocessed datasets in the following way:

- Tokenization - split of texts into tokens (in our case words),
- Removing stopwords and words with a length smaller than three characters,
- Word lemmatization and normalization to lowercase form,
- Selection of words which occurred in at least ten documents, but in less than 50% of documents.

After preprocessing, we took 2000 most frequent words.

For the evaluation of our experiments, we decided to select three standard evaluation metrics.

First, we used UMass topic coherence [20] to evaluate the quality of the extracted topics. It represents the pairwise score of *n* top words of the topic and is defined as follows:

$$coherence = \sum_{i<j} log \frac{D(w_i,w_j)+1}{D(w_i)} \quad (1)$$

where $D(w_i)$ is defined as a number of documents containing the word w_i and $D(w_i, w_j)$ is a number of documents containing both words w_i and w_j.

As a second metric we use normalized pointwise mutual information (NPMI) topic coherence [21]:

$$NPMI(w_i) = \sum_{j}^{N-1} \frac{log\frac{P(w_i,w_j)}{P(w_i)P(,w_j)}}{-logP(w_i,w_j)} \quad (2)$$

As a last, we applied normalized mutual information (NMI) [22], which evaluates how diverse the topics are. For NMI we at first selected top N words (in our case N = 100)

[1] http://qwone.com/~jason/20Newsgroups/

[2] https://martin-thoma.com/nlp-reuters/

for each topic and divided them into 10 clusters {<w_1:w_{10}>,<w_{11}:w_{20}>,...,<w_{91}:w_{100}>}. Then, we used NMI metric to compare difference of those clusters between each two generated topics.

Fig. 1 and Fig. 2 presents the results of average topics coherences for the compared methods using different numbers of topics on the 20Newsgroup dataset. As we can see from the graph, LDA outperforms LSI according to UMass and also NPMI coherence evaluation metrics.

In Fig. 3 is shown that only in NMI metric LSI outperforms LSA, which means LSI generate topic with more different words. We were not able to compare HDP model using different numbers of topics, because HDP estimate number of topics on its own, so for 20Newsgroup dataset it finds 150 topics and average UMass = -295, average NPMI = 0.157, average NMI = 0.266, which are values more likely to LDA model.

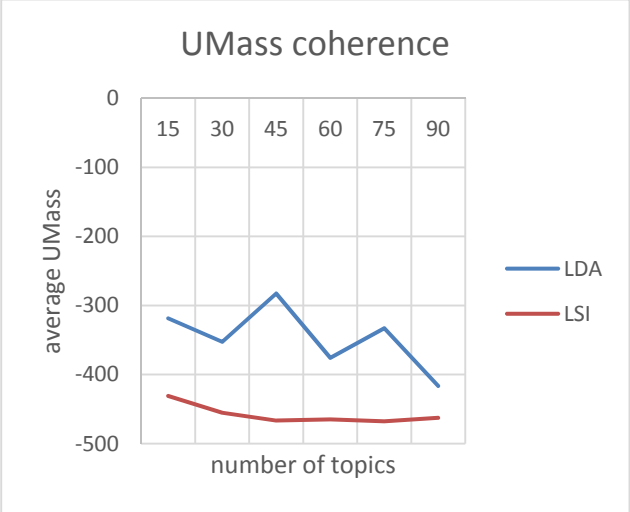

Fig. 1 Average topic UMass coherence for LDA, LSI on 20Newsgroups dataset (higher value is better)

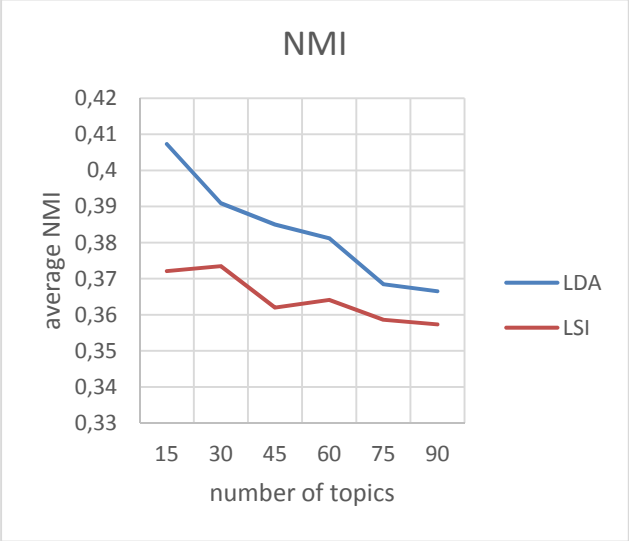

Fig. 3 Average NMI for LDA, LSI on 20Newsgroups dataset (lower value is better)

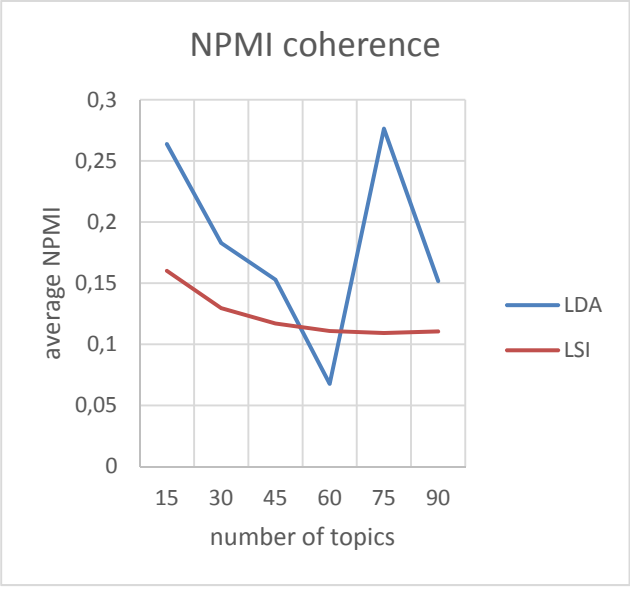

Fig. 2 Average topic NPMI coherence for LDA, LSI on 20Newsgroups dataset (higher value is better)

Fig. 4 and Fig. 5 present result of average topics coherences on Reuters dataset. From figures, you can see that LDA get slightly better results for each of coherence metrics. However similar as for 20Newsgroup dataset it achieves worse result using NMI metrics (Fig. 7).

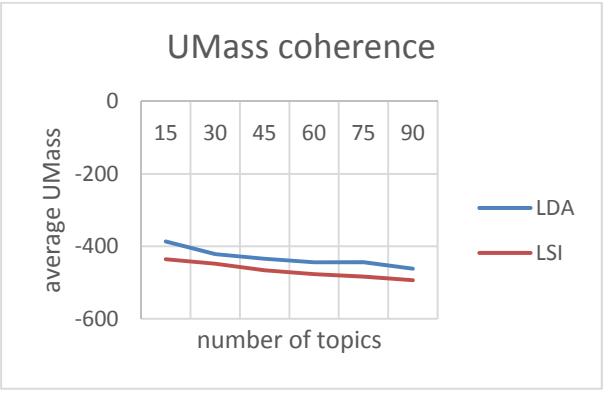

Fig. 4 Average topic UMass coherence for LDA, LSI on Reuters dataset (higher value is better)

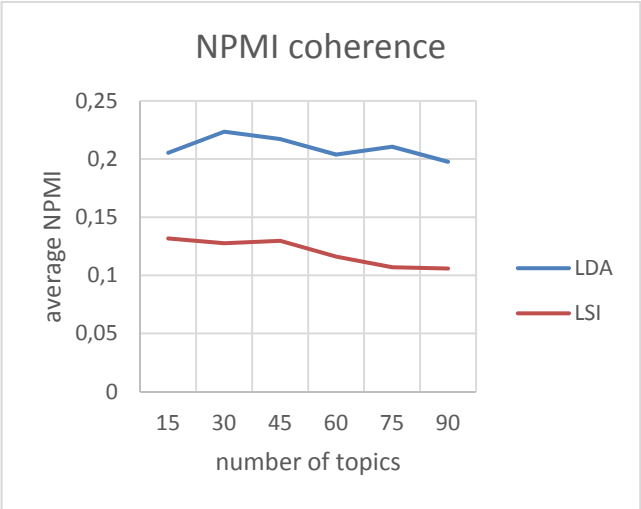

Fig. 5 Average topic NPMI coherence for LDA, LSI on Reuters dataset (higher value is better)

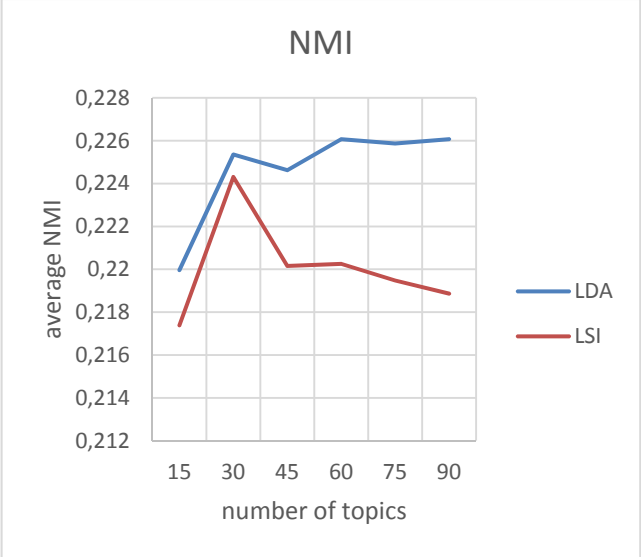

Fig. 7 Average NMI for LDA, LSI on Reuters dataset (lower value is better)

Using HDP model on Reuters dataset we find 150 topics and average UMass = -632, average NPMI = 0.07, average NMI = 0.246.

4. PROPOSED VISUALIZATION TOOL

In this section, we propose a tool for an easy and intuitive way to process, analyze and visualize textual documents. Based on the experiments from the previous section we decided to use LDA model as a method to analyze those data.

The proposed web application is used for text documents analysis and visualization. In the pilot phase, we extracted contributions from Twitter, but the user can upload corpus with his texts.

Fig. 8 Example of the search screen

Analysis by proposed tool includes topics extraction, tracking of topic evolution in time, keywords extraction, as well as the extraction of most informative texts and sentiment analysis (if the user has positive or negative attitude to current contribution). Using an appropriate visualization of acquired data we try to provide the tool that could be used by companies for decision making support. The advantage of the proposed system is fully automatic

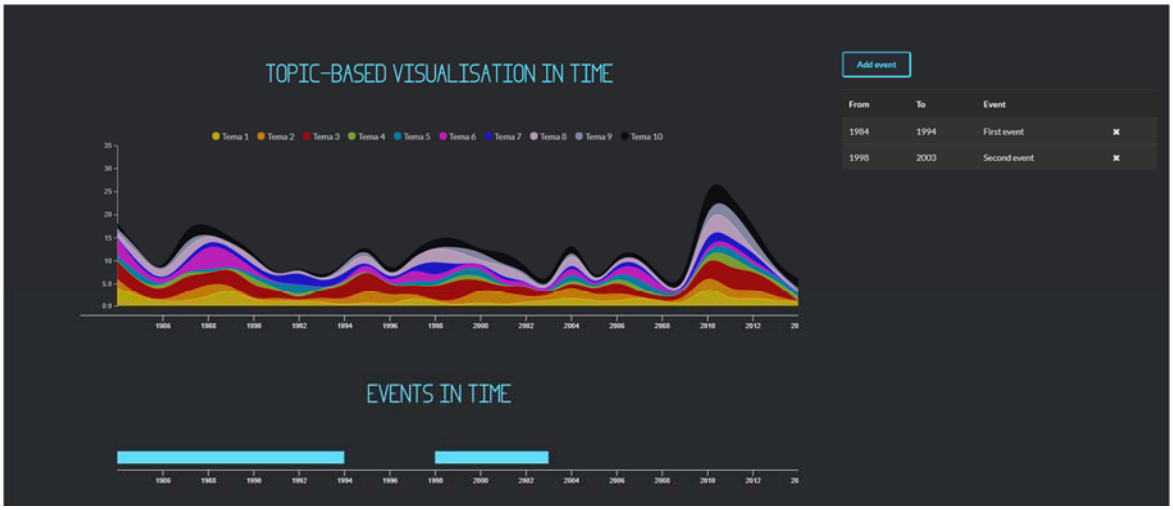

Fig. 6 Screen with topic modelling results

text analysis. The user only set areas of interest, from which text is extracted and processed.

We introduce the functionality of this tool by simple use-case. We can imagine a company, which the main scope is to selling mobile phones and with so many social networks and online data they want to use them in their favor.

Fig. 9 Example of topic detail screen

First, when the company wants to use the proposed tool, they need to upload their data to it. Next step is to search in those data by using simple query (e.g., iPhone) or more advanced search by date range and so on. Example of a search screen is shown in Fig.7. After that user sees the result of topic modeling analysis over query documents shown in Fig.8. The upper part of screen shows evolution of topic over time and in the bottom part of the screen can user define events (e.g., when was phone launched, marketing campaign, etc.) and see if there is some correlation between events and topics. Also, after clicking on some topic user is redirected to screen with a detail view of the selected topic. Example of such screen is shown in Fig.9., where the user can see topic keywords, most informative documents or contributions, and topic sentiment evolution over time.

5. CONCLUSIONS

In. this paper we presented a survey of topic modeling methods and evaluated the quality of selected ones. From experiments is clear that LDA outperform other methods. We also propose a tool for analysis of textual data and visualization of its results. Usage of this tool was presented in a simple example. We assume that this tool is potentially useful in companies as a decisions support system. One from the usage of this tool can be elections where it can track topics for each of the candidates. In the future we want to add automatic crawling of documents from web and also add more analytics methods not only topic modelling and sentiment analysis.

ACKNOWLEDGMENTS

This work was supported by the Slovak Research and Development Agency under contract No. APVV-16- 0213, under contract No. APVV-17-0267 and under contract No. APVV-SK-AT-2017-0021. This work was also supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/0493/16.

REFERENCES

[1] LANDAUER, T. K. – FOLTZ, P. W. – LAHAM, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

[2] BLEI, D. M. – NG, A. Y. – JORDAN, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[3] PETTERSON, J. – BUNTINE, W. – NARAYANAMURTHY, S. M. – CAETANO, T. S. – SMOLA, A. J. (2010). Word features for latent dirichlet allocation. In Advances in Neural Information Processing Systems (pp. 1921- 1929).

[4] ZHAI, K. – BOYD-GRABER, J. (2013). Online Latent {D} irichlet Allocation with Infinite Vocabulary. In Proceedings of the 30th International Conference on Machine Learning (pp. 561-569).

[5] TEH, Y. W. – JORDAN, M. I. – BEAL, M. J. – BLEI, D. M. (2012). Hierarchical dirichlet processes. Journal of the american statistical association.

[6] LI, X. M. – OUYANG, J. H. – LU, Y. (2015). Topic modeling for large-scale text data. Frontiers of Information Technology & Electronic Engineering, 16, 457-465.

[7] HOFFMAN, M. D. – BLEI, D. M. – WANG, C. – PAISLEY, J. (2013). Stochastic variational inference. The Journal of Machine Learning Research, 14(1), 1303-1347.

[8] PHAN, X. H. – NGUYEN, L. M. – HORIGUCHI, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web (pp. 91-100). ACM.

[9] SRIDHAR, V. K. R. (2015, June). Unsupervised topic modeling for short texts using distributed representations of words. In Proceedings of NAACL-HLT (pp. 192-200).

[10] QUAN, X. – KIT, C. – GE, Y. – PAN, S. J. (2015, June). Short and sparse text topic modeling via self-aggregation. In Proceedings of the 24th International Conference on Artificial Intelligence (pp. 2270-2276). AAAI Press.

[11] BLEI, D. M. – LAFFERTY, J. D.: Dynamic topic models, *ICML*, 2006, pp. 113-120.

[12] WANG, CH. – BLEI, D. – HECKERMAN, D.: Continuous time dynamic topic models, *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 579-586.

[13] BEYKIKHOSHK, A. *et al.*: Discovering topic structures of a temporally evolving document corpus, *KAIS*, vol 55, pp. 599-632, 2018.

[14] BLEI, D. M. – GRIFFITHS, T. L. – JORDAN, M. I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies, *JACM*, Vol. 57, 2010.

[15] HOFFMAN, T.: The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data, *IJCAI*, Vol. 99, 1999.

[16] SMITH, A. – HAWES, T. – MYERS, M.: Hierarchie: Interactive Visualization for Hierarchical Topic Models, *ILLVI*, 2014, pp. 71-78.

[17] LI, W. – MCCALLUM, A.: Pachinko allocation: Scalable mixture models of topic correlations, *JMLR*, 2008, Submitted.

[18] CHEN, P. *et al.*: Progressive EM for Latent Tree Models and Hierarchical Topic Detection, *AAAI*, 2016, pp. 1498-1504.

[19] PAISLEY, J. *et al.*: Nested hierarchical Dirichlet processes, *TPAMI*, Vol. 37, pp. 256-270, 2015.

[20] MIMNO, D. *et al.*: Optimizing semantic coherence in topic models, *Proceeding of EMNLP* 2011, pp. 262-272, 2011.

[21] SRIVASTAVA, A. – SUTTON, CH.: Autoencoding Variational Inference For Topic Models, *ICLR*, 2017.

[22] MCDAID, A. F. *et al.*: Normalized Mutual Information to evaluate overlapping community finding algorithms, 2011.

Received March 6, 2019, accepted April 9, 2019

BIOGRAPHIES

Miroslav Smatana was born in Poprad, Slovakia in 1991. He received the B.S. and M.S. degrees in artificial intelligence from the Technical University of Kosice in 2015. He is currently pursuing a PhD. degree in business intelligence at the Technical University of Kosice. His research interest includes natural language processing (especially topic modeling), data stream processing, deep learning, big data technologies, and his dissertation thesis topic is about using methods of conceptual analysis in data streams.

Viktória Martinková is MSc student at the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics at Technical University in Košice. Her study is in the branch of business information systems, and the master thesis is related to the comparison of topic modeling methods.

Dominika Maršáleková received her MSc degree in 2018 in business information systems at the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics at Technical University in Košice. Her master thesis was related to visualization of topic models and their changes (evolution) in time.

Peter Butka received his MSc degree in 2003 and PhD degree in 2010, both at the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics at Technical University in Košice. From 2014 he is an associate professor at the same department in the branch of business information systems. His research interests include data/text mining, knowledge management, semantic technologies, and information retrieval.